Title

Bayesian Meta-analysis of Multiple Continuous Treatments with Individual Participant-Level Data: An Application to Antipsychotic Drugs.

Journal

Medical Decision Making, 39(5)

Authors

Spertus, Jacob

Horvitz-Lennon, Marcela

Normand, Sharon-Lise

Publication Date

2019-07-01

DOI

10.1177/0272989X19856884

Peer reviewed

HHS Public Access

Author manuscript

Med Decis Making. Author manuscript; available in PMC 2020 August 02.

Published in final edited form as:

Med Decis Making. 2019 July ; 39(5): 583–592. doi:10.1177/0272989X19856884.

Bayesian Meta-Analysis of Multiple Continuous Treatments with Individual Participant-Level Data: An Application to Antipsychotic Drugs

Jacob Spertus[1], **Marcela Horvitz-Lennon**[2,3], **Sharon-Lise T. Normand**[1,4]

[1]Department of Health Care Policy, Harvard Medical School, Boston, MA

[2]Cambridge Health Alliance, Harvard Medical School, Cambridge, MA

[3]RAND Corporation, Santa Monica, CA

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

Abstract

Modeling dose-response relationships of drugs is essential to understanding their safety effects on patients under realistic circumstances. While intention-to-treat analyses of clinical trials provide the effect of assignment to a particular drug and dose, they do not capture observed exposure after factoring in non-adherence and dropout. We develop a Bayesian method to flexibly model dose-response relationships of binary outcomes with continuous treatment, permitting multiple evidence sources, treatment effect heterogeneity, and non-linear dose-response curves. In an application, we examine the risk of excessive weight gain for patients with schizophrenia treated with the second generation antipsychotics paliperidone, risperidone, or olanzapine in 14 clinical trials. We define exposure as total cumulative dose (daily dose×duration) and convert to units equivalent to 100mg of olanzapine (OLZ doses). Averaging over the sample population of 5891 subjects, median dose ranged from 0 (placebo randomized participants) to 6.4 OLZ doses (paliperidone randomized participants). We found paliperidone to be least likely to cause excessive weight gain across a range of doses. Compared to 0 OLZ doses, at 5.0 OLZ doses, olanzapine subjects had a 15.6% (95% CrI: 6.7, 27.1) excess risk of weight gain; corresponding estimates for paliperidone and risperidone were 3.2% (1.5, 5.2) and 14.9% (0.0, 38.7) respectively. Moreover, compared to non-black participants, black participants were associated with a 6.8% (1.0, 12.4) greater risk of excessive weight gain at 10.0 OLZ doses of paliperidone. Nevertheless, our findings suggest that

Conflict of Interest

The Authors declare that there is no conflict of interest.

Data

This study, carried out under YODA Project No. 2015-0678, used data obtained from the Yale University Open Data Access Project, which has an agreement with Janssen Research & Development, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or Janssen Research & Development, L.L.C. The CATIE data used in this paper reside in the NIH-supported NIMH Data Repositories [NIMH Data Repositories DOI: 10.15154/1373363]; Principal Investigators of original data: J. Lieberman (N01-MH090001) and P. Sullivan (R01-MH074027).

The CATIE data is available from the National Institute of Medical Health Repository and Genomics Resource (https://bioq.nimhgenetics.org/studies/?studyId=20). The Janssen Trials are available from the Yale Open Data Access Project (http://yoda.yale.edu/multiple-ncts-optics-trial-bundle).

paliperidone is safer in terms of weight gain risk than risperidone or olanzapine for all participants at low to moderate cumulative OLZ doses.

Introduction

Patients, clinicians, and other decision-makers need to know which treatments are most effective and safe. For pharmacological treatments, a key feature potentially affecting both is drug dosing. For example, while dose-finding studies provide valuable information in early phase trials, in later phase randomized controlled trials (RCTs), patients may not fully comply with their treatment assignments or remain on treatment throughout the trial. Moreover, most trials are insufficiently powered and of insufficient duration to assess a treatment dose-response relationship with safety outcomes. Antipsychotic drug trials for the treatment of schizophrenia are no exception. Because antipsychotic drugs are associated with significant metabolic risk and most patients with schizophrenia must remain on these drugs for a lifetime, characterizing the relationship of observed total cumulative dose, defined as daily dose multiplied by duration on treatment, with meaningful safety outcomes is critical[1-4]. Additionally, understanding if and how patient characteristics modify efficacy or safety outcomes, a phenomenon known as treatment effect heterogeneity (TEH), is necessary for optimal treatment decisions in diverse patient populations.

We present a case study of methods to estimate dose-response functions and dose-dependent TEH using clinical trial data, which we motivate with an application in psychiatry. We use individual participant-level data (IPD) from more than a dozen RCTs assessing outcomes of specific antipsychotic drugs. Our safety outcome is excessive weight gain operationalized as a binary variable assuming a value of 1 if the participant gained at least 7% of his/her baseline weight. The binary endpoint was selected rather than a continuous weight gain outcome because the binary outcome captures the rate of clinically relevant weight gain, and it is widely used in pharmaceutical trials and is therefore interpretable and meaningful for prescribers, researchers, and regulators[3,5]. Evidence on the association between dose and weight gain for people exposed to antipsychotics is inconsistent, and methodological considerations limit its value[6]. Due to the limited and mixed evidence of TEH of antipsychotic safety effects, we are also interested in assessing whether race/ethnicity moderates the weight effects of antipyschotics and whether this moderating effect varies by dose[7,8].

The RCT data provide key variables on dosing, outcomes, compliance, and potential confounders, and have the advantage of randomization within each trial so that, on average, participant measured and unmeasured characteristics should be balanced among treatment arms. However, there are at least 4 major challenges to dose-response function estimation in this setting. First, actual total cumulative dose taken is not randomized and is therefore confounded by baseline variables. Second, because TEH is crucial to providing guidance for decision makers dealing with diverse populations and single trials are rarely powered to detect TEH, combining data from multiple trials to characterize dose-response relationships is necessary. Third, dose-response relationships are likely to be non-linear and necessitate the application and interpretation of flexible regression techniques. Fourth, although pooling

information from multiple RCTs is essential, meta-analysis is complicated in this setting because interest centers on curves (e.g., dose-response) characterized by multiple parameters rather than means (e.g., average outcome at one specific dose).

We meet these challenges through the use of a Bayesian hierarchical model as this naturally incorporates multiple levels of heterogeneity, while posterior parameters of interest can be summarized flexibly and intuitively. We adjust for confounding by including controls in an outcome model and baseline patient covariates. We permit estimation of flexible dose-response functions through the use of splines, where the coefficients are distributed hierarchically to partially pool information across trials. Curves can be estimated for multiple distinct treatment drugs. By combining evidence across trials and using flexible hierarchical splines to model the dose-response relationship, we provide an important approach to assessing dose-response functions. We further define TEH for a binary moderator, allowing separate curves to be estimated for two subgroups. We introduce and discuss some inferential tools that can provide powerful and intuitive interpretation of results in this complex setting. To our knowledge, there has been limited discussion of Bayesian approaches in the multiple continuous treatment setting, a gap we aim to fill in this paper.

Methods

YODA and CATIE Trials

To illustrate methods, we estimate dose-response curves describing second generation antipsychotics' (SGA) effect on weight gain by combining evidence from multiple individual participant-level clinical trial datasets. The model to estimate the curves is described below. We use data from 13 clinical trials obtained from the Yale Open Data Access (YODA) project and supplement them with data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) trial. These trials were designed to assess efficacy/effectiveness of antipsychotic in people with schizophrenia, primarily on the basis of the Positive and Negative Syndrome Scale (PANSS) score, a widely used measure of severity of schizophrenia symptoms, but also assessed weight gain and other safety outcomes. Trials varied in duration, from 6-weeks (six YODA trials) to 18-months (CATIE). We include adult patients randomized to 1 of 3 SGAs or to placebo (N= 5891) (see Table 1). The outcome variable is a binary indicator of whether a participant gained more than 7% of their baseline body weight. We calculate total cumulative dose by summing prescribed daily doses over the duration of treatment (accounting for dropout). Adherence data in the form of pill counts or injections are also incorporated when available. For pill trials, even those with adherence data available, there is inherently some uncertainty in exposure because subjects were not always monitored and may not have taken all of their assigned dose. Injection trials have no exposure uncertainty as treatment was fully administered at trial centers. In a prior study, we found no substantial differences due to the form of treatment and do not account for such differences here[9].

Because the potency of different SGAs varies, we place them on the same scale by converting to doses equivalent to 100mg of olanzapine (OLZ doses), a scaling approach in psychopharmacology supported by empirical research[10-12]. We include information from placebo arms to permit more precise estimation of the intercepts and coefficients for

confounders. As baseline confounders, we include age, race, sex, body mass index (BMI), and the total PANSS score. These clinically relevant variables were available in all trials (Table 1). Further details of our selection and handling of these trials can be found in Spertus et al (2018)[9].

Causal Assumptions

Because total cumulative dose is not randomized, we make some additional assumptions for causal inference. Imai and van Dyke detailed these assumptions in the context of a (single) continuous treatment regime[13]. For our results to be interpreted causally, these assumptions must hold both within and between trials. First, we assume the stable unit treatment value assumption (SUTVA) holds in that total cumulative dose for a given subject does not affect weight gain for any other subject. Second, we assume strong ignorability such that the potential outcomes are independent of treatment received conditional on measured confounders. Random assignment of drugs within trials helps satisfy this assumption. However, cumulative dose taken was not randomized and we rely on adjusting for measured confounders to satisfy this assumption. Finally, we assume positivity meaning that any participant could receive any drug at any (non-negative) dose within groups of individuals defined by their confounders. We note that positivity with a continuous exposure is not empirically verifiable, that is, if treatment is modeled as a continuous variable, it is generally impossible for two patients to have the same dose. However, because trial participants were randomized to drug doses and had similar inclusion criteria, positivity is valid by design.

Model

We use a fully Bayesian hierarchical model to estimate dose-response curves for each drug while allowing heterogeneity across trials. Specifically, let $j \in \{1,\ldots J\}$ index trials, $i \in \{1, \ldots I\}$ index subjects, and $k \in \{1, \ldots K\}$ index treatments. Let Y_{ij} be a binary outcome for subject i in trial j, $\mathbf{T}_{ij} = (T_{ij1}, T_{ij2} \cdots T_{ijK})$ be a length K vector of exposure intensities with T_{ijk} the amount of drug k taken, and $\boldsymbol{X}_{ij}$ be a length P vector of potential confounders. We posit a hierarchical logistic regression of Y_{ij} on treatment and confounders:

$$Y_{ij} \sim \text{Bernoulli}\left\{\text{logit}^{-1}\left[\alpha_j + \sum_{k=1}^{K} f_{jk}(T_{ijk}) + \sum_{p=1}^{P} \beta_p X_{ijp}\right]\right\}. \quad (1)$$

α_j is a trial specific intercept and β_p is a population-level linear term for confounder p. Thus we assume that confounder effects are linear on the logit scale, additive, and the same across trials. In contrast, the dose-response functions $f_{jk}(T_{ijk})$ are modeled non-linearly and hierarchically.

We model the functions f_{jk} using B-splines with different numbers of knots, where the coefficients for each element are hierarchically modeled to pool information across trials. B-splines are piecewise polynomials in which the pieces are joined at the knots. Varying numbers of knots are tried and are always equally spaced at quantiles of each drug: a model with 0-knots is linear in all groups; a 1-knot model is a B-spline with a knot at the median of each drug dose; a 2-knot model has knots at the .33 and .66 quantile of each drug dose; and

so on[14]. The number of knots in the final model is chosen based on leave-one-out information criterion, a fully Bayesian estimate of out-of-sample fit[15].

To complete the specification, we place weakly informative priors on every parameter following best practices in Bayesian data analysis[16]. Further details are available in the appendix (section A).

Treatment Effect Heterogeneity

We expand on the model above in order to account for TEH, i.e. different dose-response curves for subgroups of the patient population. We introduce a binary moderator variable to investigate potential TEH between two subgroups, although this could be readily expanded to more subgroups. Briefly, this involves further splitting the functions f_{jk} so that separate curves are defined for each value of the binary moderator. We do this by introducing separate coefficients for each level of the moderator, without changing how the underlying B-splines are defined. Priors are as above. Interpretation is complicated when there is TEH, and we discuss approaches below. For further details on defining a model with TEH see the appendix (section B).

Model Fitting and Checking

We use Monte Carlo sampling to estimate posterior quantities of interest implemented via Hamiltonian Monte Carlo in Stan that typically requires only a few thousand draws[17]. We use the R wrapper package brms to fit the models[18]. The quality of the draws is assessed by examining trace plots for a few quantities (particularly the log-posterior) to identify any pathologies. Convergence is summarized by the Gelman-Rubin R-hat statistic and declared achieved when all parameters have an R-hat below 1.1[16].

To check how well the model captures elements of the data, we examine posterior predictive distributions of marginal outcome rates within treatment groups, as well as the maximum and minimum marginal outcome rates across trials. These are then compared with the corresponding rates observed in the actual trials. If the model fits well, the observed rates should fall well within the distribution of posterior predicted rates. Such posterior predictive checks are a standard model diagnostic in Bayesian inference[16].

Interpretation and Inference

By using a Bayesian approach that returns the posterior distribution for all parameters, inference is readily facilitated as appropriate functions of the posterior. However, summarizing the posterior in a useful and interpretable way for decision making is non-trivial because the dose-response effects are non-linear curves defined by multiple parameters. We briefly outline these approaches below, for more detail see the appendix (section C).

First, we provide plots of the marginal dose-response curves, along with 95% credible intervals. Plotting the curves provides the most obvious and intuitive way to interpret the posterior. A decision maker can readily see the probability of an adverse outcome over a range of doses for specific drugs, providing key evidence to select the most appropriate drug

and dose for a given subject. Second, for each drug k we compute the average treatment effect at fixed dose a, defined as the expected outcome probability for a subject receiving dose a of drug k minus the expected outcome probability if they had received 0 dose. This is averaged over subjects and over posterior draws, and uncertainty is summarized by 95% credible intervals. Third, we compute rank probabilities, defined as the probability that drug k has the smallest effect among all drugs over a range of doses between 0 and A. The rank probability provides a single number summary of how likely a drug is to be the "best", i.e. to have the lowest effect on the rate of excessive weight gain.

When considering potential TEH, a different approach is needed. We compute and plot a function that we call the "difference curve", which summarizes the additional treatment risk for a given subgroup over a range of doses. The difference curve is defined as the difference in average treatment effects between two subgroups at each dose a between 0 and A. It can be plotted long with uncertainty intervals and examined, where a substantial difference indicates potential TEH at that dose.

Implementation

We implement the model described above for analysis of the YODA and CATIE data. A model with linear terms for treatment provides a baseline for comparison, and more complex models are built using splines with increasing numbers of knots and compared. Spline bases are generated for each treatment using the bs R function, with knots placed at appropriate percentiles. We built a model for TEH by including race as a binary moderator variable (black, non-black). We test models of increasing complexity (linear, 1 knot, 2 knots…) using the leave-one-out information criterion, and stop adding knots when the out-of-sample fit worsens.

Results

Data

Table 1 summarizes information on covariates, outcomes, and exposure. Olanzapine and risperidone are measured in fewer trials and on fewer subjects than paliperidone, which is likely to affect the precision of their estimated dose-response curves. There are some imbalances in baseline covariates that may confound unadjusted estimates. The rate of excessive weight gain varies considerably between treatments, as does the distribution of exposures, with median total cumulative dose highest for paliperidone and lowest for risperidone. In addition to the 1368 placebo subjects, 10 olanzapine, 56 paliperidone, and 22 risperidone subjects have zero exposure. Eight percent of olanzapine, 5% of paliperidone, and 19% of risperidone subjects have an exposure of less than 100mg OLZ doses, while 81% of olanzapine, 41% of paliperidone, and 78% of risperidone subjects have an exposure of less than 500mg OLZ doses.

Model

Convergence of our MCMC algorithms is indicated by trace plots and R-hat statistics. The LOO-IC suggests that a spline with a single knot provides the best fitting model. Graphical posterior predictive checks indicate that the model fits the observed data well in that it

replicates key aspects of the observed data. Observed outcome rates within treatment groups and trials, as well as the observed maximum and minimum across trials fall well within the posterior predictive distributions drawn from the model.

Inference

Table 2 displays average treatment effects of moving from zero dose (no treatment received) to dose intensities equivalent to either 100mg or 500mg of olanzapine for all three drugs studied, along with 95% credible intervals. It is immediately apparent that higher total cumulative doses increase the probability of excess weight gain for all three drugs. While olanzapine and risperidone have the largest average effects, there is considerable uncertainty in their effect sizes as well.

Treatment effects along a range of exposures from 0 to 800mg OLZ equivalents are displayed in Figure 1. Paliperidone has a small positive effect on the risk of weight gain across doses. At high total cumulative dose, the effect on weight gain becomes very uncertain for olanzapine and risperidone, likely due to the small number of subjects at those doses. The effects level off at high exposures, although the high variance in these dose regions complicates the interpretation of these results.

Figure 2 depicts paliperidone TEH over the range of exposure by subtracting the curves for non-black subjects from the curves for black subjects. If the exposure has a greater effect on weight gain for black subjects, the curve will be positive at that total cumulative dose. The lower 95% credible limit of the curve rises above 0 starting at around 700mg OLZ equivalent doses, though it dips back below again around 1200mg OLZ doses. At high doses, the credible limits broaden because there are fewer patients and the curve flattens out, possibly indicating a saturation effect. The difference curve provides evidence of potential TEH: compared with non-black subjects, black subjects are expected to have an increased weight gain rate of about 5% on average at 700mg OLZ equivalents or more of paliperidone. For the other two drugs, 95% credible limits cover 0 at all exposures, indicating no evidence of TEH, a finding likely due to smaller sample sizes for risperidone or olanzapine.

We compute the posterior probabilities that each drug is the best in terms of having the smallest effect on weight gain on average over a range of exposures, from 0 to 500mg OLZ equivalents. 500mg OLZ equivalents represents a low average daily dose over 6 weeks (about 12 mg/day, given a dose range of 10-20 mg/d), and is near the center of the range of doses observed in each treatment group in our data. Probabilities are estimated for both non-black and black subjects, but subgroups estimates are very similar over this range so we refer here to the overall results. Paliperidone has a probability of .85, risperidone has a probability of .06, and olanzapine has a probability of .09 of being the best in terms of weight effects. Thus, the evidence strongly suggests that paliperidone is least likely to cause excessive weight gain over this range of exposures.

Discussion

In this paper we introduced a framework for hierarchical modeling of non-linear treatment effects with potential TEH. Our work was motivated by the need to address methodological

shortcomings limiting the utility of RCTs to yield dose-response relationships of pharmacological treatments accounting for non-compliance, dropout, and differential dosing. We characterized the weight effects of three SGAs (olanzapine, paliperidone, and risperidone) across a range of exposure assessed through standardized total cumulative doses and assessed whether race moderated the observed effects. We chose excessive weight gain as our safety outcome because weight gain is associated with poor metabolic outcomes, including metabolic syndrome, dyslipidemia, hypertension, and type 2 diabetes, all of which are risk factors for cardiovascular disease. This relationship is particularly significant in the context of a rapid growth in SGA utilization in the U.S., partly driven by their frequent use for off-label indications[19].

We found that these commonly used SGAs led to increased probability of weight gain over their range of exposures, although olanzapine and risperidone provided very imprecise estimates of effect curves. There seemed to be a leveling off of the effects at higher doses, a feature that we were able to capture using a non-linear model. Moreover, we found evidence that race may moderate the association of paliperidone with excessive weight gain, with potentially increased risk of weight gain for black compared to non-black participants at moderate doses of paliperidone. At high doses we did not detect significant TEH due to fewer subjects and increased uncertainty in that range. Despite potentially increased risk for black participants, our results suggest that paliperidone is likely the best among the three SGAs for both black and non-black patients. We were unable to detect any moderation of race on the effects of risperidone or olanzapine due to high levels of uncertainty.

We believe that our findings are an important contribution to the evidence on the dose-dependency of safety effects of SGAs and TEH on antipsychotic effects, a body of research that has been encumbered by methodological challenges we have sought to address. In terms of dose-dependency of antipsychotic weight effects, the evidence is generally inconclusive[6]. While some studies suggest a dose-response relationship for olanzapine, the evidence is mixed for risperidone and quetiapine[20-24]. In terms of race-related TEH for weight effects of SGAs, the evidence is limited and mixed and we are not aware of studies that have assessed whether race moderation varies by dose[25,26].

Additionally, we have expanded on the statistical literature by introducing an IPD meta-analytic framework for assessing dose-response relationships from multiple continuous treatments with potential TEH. For past work in estimating dose-response curves using multiple data sources see Gasparini et al (2012) and Bobb et al (2013), both of which develop methodology for use in environmental statistics[14,27]. For a review of meta-analysis and its use in medical statistics, see Normand (1999)[28]. A recent review of TEH and its important place in causal inference for medical applications can be found in Kent et al (2018)[29]. Xie (2013) provides a more formal treatment of TEH in causal inference[30]. The Bayesian hierarchical framework we used, along with many technical points of our data analysis are covered well in Gelman et al (2014)[16].

We also focused on developing powerful ways to interpret the results of our analysis. Our work permits meaningful and actionable inferences from dose-response curves for clinicians faced with complex treatment and dosing decisions, and can also inform decisions by

administrators and regulators involved in the care of these patients. A Bayesian framework facilitated this effort as estimates and credible limits could easily be computed from the posterior draws. Plotting non-linear dose-response curves along with credible limits is a clear first step in analyzing multiple continuous treatments and can give a rough sense of adverse event probabilities over a range of doses while revealing threshold or saturation effects. Computing the average treatment effect of moving from 0 dose (off treatment) to a given dose for each drug provides a more precise numerical summary of treatment risk averaged over the patient population. For a fixed dose and drug, this estimate summarizes the likelihood that a subject may experience clinically meaningful weight gain. This information may be especially useful for shared clinician-patient decision making where trade-offs between expected treatment and efficacy and adverse effects must be weighed based on likelihood of risk and patient preferences. Rank probabilities provide a scalar summary of the safety or efficacy of a given drug over a range of doses. In this context, a decision-maker would generally want to choose the drug that has the highest rank probability, the best drug, as this drug is likely to lead to the least adverse effects over the analyzed range of doses. Such scores may be particularly useful in a policy setting context.

In order to assess possible TEH, we computed and plotted difference curves, which allow visualization of the magnitude and uncertainty of effect differences between groups hypothesized to have moderating effects over a range of exposure values. Difference curves allow decision-makers to assess whether patients with potentially effect-modifying characteristics should be given different treatments. Evidence of meaningful differences in treatment effects, whether efficacy or safety-related, should inform decision-making by patients and their families, clinicians, administrators, and regulators. As they are complex functions of the posterior, caution in interpretation is warranted and decision curves should not be treated as authoritative evidence of TEH. They provide evidence that TEH may exist and could be especially useful as exploratory tools to indicate drugs and doses that are promising for follow-up experiments. A strength of our approach is that it generalizes readily to other drug trials and applications outside psychiatry where one or more continuous treatments are of interest.

Limitations

In order to assess the rate of excessive weight gain, as opposed to average weight gain, we dichotomized our outcome into a binary indicator. Modeling a continuous outcome would likely have led to more precise estimates of dose-response curves. We chose to dichotomize the outcome to directly model the rate of excessive weight gain. However, future work could improve on our estimates by initially modeling a continuous outcome before computing the probability of a > 7% increase *post hoc,* based on the posterior. Another challenge in our application, and one that is likely to arise in other multivariate continuous treatment designs, is how to standardize exposures. We chose to use olanzapine equivalent doses, but other absolute scales such as chlorpromazine equivalents or recommended daily doses may also be used. As these standardized doses are themselves estimated, a source of uncertainty is ignored in order to make treatments comparable. Another possibility is to scale the treatments relatively, e.g. divide the treatment variables so 100 is always the maximum dose, but this could introduce problems if certain drugs were included in the original trials at

relatively higher doses. Additionally, we had few trials to assess olanzapine and risperidone (4 and 2 respectively), leading to high uncertainty in these estimates. We extracted meaningful inferences for paliperidone, but many of the results for olanzapine and risperidone were so unclear that they will not be very informative for decision makers. An exception is the rank probabilities, which provided relatively strong evidence that paliperidone was superior to both olanzapine and risperidone. Nevertheless, we demonstrated the power of our methods to make decisions when multiple drugs and doses are involved. As more individual participant level trial data is made available, it will be possible to use our framework with additional information to make more precise inferences on more drugs and better understand optimal treatments for diverse subjects. Another potential limitation is that our confounders do not include comorbidities and concurrent use of other drugs (or services) potentially associated with weight changes. Although this may have affected our results, their impact is likely to be minimal since all patients were randomized to the study drugs.

Conclusion

We believe this paper contains important scientific as well as methodological contributions. Our findings are scientifically valuable because we have expanded the evidence of dose-dependency and race-based moderation of the weight effects of SGAs, a widely used class of antipsychotics that is associated with elevated metabolic risk. Methodologically, we provided a way to draw inferences on the impacts of multiple continuous treatments measured in multiple trials with potential TEH. In addition to describing a Bayesian model to accommodate these features, we developed a number of ways to summarize the posterior distribution that can lead to relevant and interpretable insights for decision-makers.

Acknowledgments

This study was approved by the IRB of the Harvard Faculty of Medicine. We would like to thank our collaborators at the Harvard Catalyst Reactor Program, Hardeep Ranu and Gary Gray; Marsh Wilcox, Janssen Scientific Director & Fellow; Linda Valeri, Harvard Medical School; and John Jackson, Johns Hopkins Bloomberg School of Public Health. We would also like to thank Katya Zelevinksy for help with data analysis and Haley Abing for assistance with data handling and management.

Funding

Mr. Spertus' effort was conducted with support from The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR001102), Harvard Catalyst, and Harvard University. Drs. Horvitz-Lennon and Normand were funded by Harvard Catalyst, the National Institute of Mental Health (R01-MH106682), and the National Institute of Minority Health and Health Disparities (R01-MDO12428). Dr. Normand was also supported by the National Institute of General Medical Sciences (R01-GM111339). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic healthcare centers, or the National Institutes of Health.

Appendix

A Model Details

Under the B-spline formulation for f_{jk}, the exposures for each drug k are expanded into recursively defined power bases with local support[1]. The bases are defined by boundary

points and fixed knots placed within the range treatment. After expansion into a dimension-L B-spline basis with elements $\eta_k(\boldsymbol{T}_k)$ we express each treatment function as a weighted sum of its basis:

$$f_{jk}(T_{ijk}) = \phi_{jk1} T_{ijk} + \sum_{l=2}^{L} \phi_{jkl} \eta_{kl}(T_{ijk}). \quad (1)$$

Equation (1) asserts that all between trial differences in treatment are expressed through $\boldsymbol{\phi}_{jkl}$, after adjusting for participant confounders.

To complete the model, we specify priors on the confounder coefficients β_p and hierarchical priors on the intercept and coefficients of f_{jk} in Equation (1). For the confounder coefficients, β_p, we simply specify the weakly informative prior: $\beta_p \sim t_5(0, 2.5)$, which places most of the prior mass in the interval [−5,5]. This is a reasonable range on the log-odds scale when confounders are binary or scaled to unit variance[2]. We let $\boldsymbol{\phi_j} = (\boldsymbol{\phi}_{j1} \ldots \boldsymbol{\phi}_{jK}) = (\boldsymbol{\phi}_{j11}, \ldots \boldsymbol{\phi}_{j1L}, \ldots \boldsymbol{\phi}_{jK1} \ldots \boldsymbol{\phi}_{jKL})$, denote the length $K \times L$ concatenation of all spline coefficients for each treatment function in trial j; the overall mean across trials of the intercept be the scalar μ_α; and the mean treatment coefficients be $\boldsymbol{\mu}_{\boldsymbol{\phi}} = (\mu_{11}, \ldots \mu_{1L}, \ldots \mu_{K1} \ldots \mu_{KL})$. The hierarchical prior is specified as:

$$\begin{bmatrix} \alpha_j \\ \boldsymbol{\phi}_j^{\mathrm{T}} \end{bmatrix} \overset{\text{iid}}{\sim} \mathcal{N}_{KL+1} \left\{ \begin{bmatrix} \mu_\alpha \\ \boldsymbol{\mu}_{\boldsymbol{\phi}}^{\mathrm{T}} \end{bmatrix}, \boldsymbol{\Sigma} \right\} \quad \text{where } \boldsymbol{\Sigma} = \mathrm{D}(\boldsymbol{\sigma}) \boldsymbol{\Omega} \mathrm{D}(\boldsymbol{\sigma}). \quad (2)$$

Here the $(KL + 1) \times (KL + 1)$ covariance matrix Σ is decomposed into the product of a diagonal matrix, $\mathrm{D}(\boldsymbol{\sigma})$, defined by a $(KL + 1)$ vector of standard deviations $\boldsymbol{\sigma}$, and a $(KL + 1) \times (KL + 1)$ correlation matrix $\boldsymbol{\Omega}$. The entries of $\boldsymbol{\sigma} = (\sigma_\alpha, \sigma_{11}, \ldots \sigma_{1L}, \ldots \sigma_{K1}, \ldots \sigma_{KL})$ represent the between trial variability of each hierarchically specified parameter.

Finally, the prior distributions for the hierarchical parameters are specified as:

$$\mu_\alpha \sim \mathcal{N}(0, \infty);\ \mu_{\phi kl} \sim t_5(0, 2.5);\ \sigma_{kl} \sim \mathrm{C}^+(0, 0.1);\ \text{and } \boldsymbol{\Omega} \sim \mathrm{LKJ}(3). \quad (3)$$

The prior on the intercept μ_α, is completely non-informative. The $t_5(0, 2.5)$ priors for $\mu_{\phi kl}$ are made to be weakly informative given the binary outcome and the scaling of the exposure (see Table 1 of the main paper)[2]. The half-Cauchy prior on σ_{kl} is slightly regularizing in the sense that it pulls σ_{kl} towards zero and thus the trial specific estimates towards their group means, μ_α or $\mu_{\phi kl}$[3]. With $\mathrm{C}^+(0, 0.1)$ about 95% of the prior mass is below 1.3, which pools the estimates when there is little information at the trial level but permits substantial variation if the data are informative. The correlation matrix, $\boldsymbol{\Omega}$, is *a priori* distributed according to a Lewandowski, Kurowicka, Joe (LKJ) distribution, determined by a single hyperparameter, which we set equal to 3 to put slightly more prior weight on the identity

matrix[4,5]. This leads to a regularized covariance matrix, $\boldsymbol{\Sigma}$, which is necessary when sparse data inform the elements of $\boldsymbol{\phi}_j$, for example, if all drugs are not studied in all trials.

B Treatment Effect Heterogeneity Details

We accommodate treatment effect heterogeneity by estimating separate curves within patient subgroups. Specifically, let the moderators $\mathbf{M} \subset \mathbf{X}$ denote a subset of the covariates that may modify the effect of dose on outcome. In our study, we posit that $\mathbf{M} = \boldsymbol{M}$ is a single binary moderator indicating if the participant is black versus non-black. The additive treatment effects within levels of M_{ij} are of primary interest. We introduce a new parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_K) = (\theta_{11}, \ldots \theta_{1L}, \ldots \theta_{K1} \ldots \theta_{KL})$, which parameterizes additional dose-response functions for each treatment for subjects with $M_{ij} = 1$, fixed across trials. The models for curves within moderator levels are:

$$f_{jk}(T_{ijk}, M_{ij} = 0) = \phi_{jk1} T_{ijk} + \sum_{l=2}^{L} \phi_{jkl} \eta_{kl}(T_{ijk}) \quad (4)$$

$$f_{jk}(T_{ijk}, M_{ij} = 1) = \phi_{jk1} T_{ijk} + \theta_{k1} T_{ijk} + \sum_{l=2}^{L} \{\phi_{jkl} \eta_{kl}(T_{ijk}) + \theta_{kl} \eta_{kl}(T_{ijk})\}. \quad (5)$$

As above, the prior on θ_{kl} is %(0, 2.5). Note that $\boldsymbol{\theta}$ does not vary by trial, which implies that the moderating effect is constant across trials so that all trial level heterogeneity is still expressed through $\boldsymbol{\phi}_j$. We also assume the basis coordinates $\eta_{kl}(T_{ijk})$ are the same for both subgroups.

For conciseness, we denote the set of all parameters as $\boldsymbol{\Upsilon}$. Draws from the posterior are indexed by $q = \{1, \ldots, Q\}$ and $\boldsymbol{\Upsilon}_q$ indicates a single draw from the joint distribution of all parameters.

C Interpretation and Inference Details

We define several approaches to assess differences in drugs across doses, drug types, and potential moderators after model fitting is complete. We compute point-wise 95% credible intervals to express uncertainty. First, we determine the dose-response curve for the "typical" subject by setting the confounders to 0. The posterior mean curve is plotted along with 95% credible intervals. We also determine the average treatment effect at fixed doses. We consider the risk difference, $\Delta(k, a)$ which is defined as the expected outcome if all subjects were treated with treatment k at dose a minus the expected outcome if all subjects were treated at no dose. A distribution on the risk difference is obtained by computing it for each draw from the posterior:

$$\Delta_q(k, a) = \mathbb{E}_{\mathbf{X}}\{\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = a, \mathbf{X}, \boldsymbol{\Upsilon}_q) - \mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = 0, \mathbf{X}, \boldsymbol{\Upsilon}_q)\}.$$

The posterior mean and credible intervals for $\Delta(k, a)$ are computed from the Q draws obtained. We summarize comparisons by integrating over a range of dose intensities and averaging over draws. For example, we calculate the probability that drug 1 has the smallest effect among all drugs over a range of intensities $a \in (0, A)$:

$$\mathbb{P}\{\Delta(k=1) = \min_k \Delta(k)\} = \frac{1}{Q}\sum_{q=1}^{Q}\int_0^A I\{\Delta_q(k=1,a) = \min_k \Delta_q(k,a)\}\,\mathrm{da}, . \quad (6)$$

The integral can be computed numerically by making a discrete mesh over the range of doses $(0, A)$. This quantity may be useful as a scalar summary of the treatment effects that can be used for decision making. It is straightforward to compute the probability that a drug has the greatest effect merely by taking the maximum instead of the minimum in (6).

A different approach is needed for interpreting treatment effect heterogeneity. We estimate subgroup-specific effects by taking local average treatment effects across subgroups. Letting $\boldsymbol{M}$ be a binary moderator of interest, we let $\Delta_q(\boldsymbol{M} = 1, \boldsymbol{T}_k = a)$ represent the risk difference in stratum $M = 1$ at dose a of treatment $\boldsymbol{T}_k$, and calculate the distributions within subgroups as:

$$\Delta_q(\boldsymbol{M}=1, \boldsymbol{T}_k = a) = \mathbb{E}_{\mathbf{X}}\{\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = a, \boldsymbol{M}=1, \mathbf{X}, \boldsymbol{\Upsilon}_q) - \mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = 0, \boldsymbol{M}=1, \mathbf{X}, \boldsymbol{\Upsilon}_q)\} \quad (7)$$

$$\Delta_q(\boldsymbol{M}=0, \boldsymbol{T}_k = a) = \mathbb{E}_{\mathbf{X}}\{\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = a, \boldsymbol{M}=0, \mathbf{X}, \boldsymbol{\Upsilon}_q) - \mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{T}_k = 0, \boldsymbol{M}=0, \mathbf{X}, \boldsymbol{\Upsilon}_q)\}\,. \quad (8)$$

Comparisons within drugs but across subgroups are then facilitated by appropriate functions of $\Delta_q(\boldsymbol{M} = 1, \boldsymbol{T}_k = a)$ and $\Delta_q(\boldsymbol{M} = 0, \boldsymbol{T}_k = a)$. A particularly useful function is the difference curve which we define as: $\Delta_q(\boldsymbol{M} = 1, \boldsymbol{T}_k = a) - \Delta_q(\boldsymbol{M} = 0, \boldsymbol{T}_k = a)$. We simplify computation of the difference curve by ignoring the covariates $\mathbf{X}$ which change the intercept. This has the effect of setting all covariates to their default values (assuming centered covariates) and allows computation of a single representative curve. The posterior mean and credible intervals are defined at each dose a, and the entire difference curve is graphed. TEH can be assessed by examining the plotted difference curve, where deviation from 0 indicates there may be TEH at that dose.

References

[1]. Nielsen J, Skadhede S, and Correll C Antipsychotics associated with the development of type 2 diabetes in antipsychotic-naive schizophrenia patients. Neuropsychopharmacology 2010;35:1997–2004. [PubMed: 20520598]

[2]. Hert MD, Schreurs V, Sweers K, et al. Typical and atypical antipsychotics differentially affect long-term incidence rates of the metabolic syndrome in first-episode patients with schizophrenia: A retrospective chart review. Schizophrenia Research 2008;101:295–303. [PubMed: 18299188]

[3]. Newcomer J and Haupt D The metabolic effects of antipsychotic medications. Can J Psychiatry 2006;51:480–491. [PubMed: 16933585]

[4]. Rummel-Kluge C, Komossa K, Schwarz S, et al. Head-to-head comparisons of metabolic side effects of second generation antipsychotics in the treatment of schizophrenia: A systematic review and meta-analysis. Schizophrenia Research 2010;123:225–233. [PubMed: 20692814]

[5]. Hert MD, Yu W, Detraux J, Sweers K, van Winkel R, and Correll C. Body weight and metabolic adverse effects of asenapine, iloperidone, lurasidone, and paliperidone in the treatment of schizophrenia and bipolar disorder. CNS Drug 2012;26:733–759.

[6]. Wirshing D, Wirshing W, Kysar L, et al. Novel antipyschotics: comparison of weight gain liabilities. Journal of Clinical Pyschiatry 1999;60.

[7]. Meyer J, Davis V, Goff D, et al. Change in Metabolic Syndrome Parameters with Antipsychotic Treatment in the Schizophrenia Trial: Prospective Data from Phase 1. Schizophrenia Research 2008;101:273–286. [PubMed: 18258416]

[8]. Patel JF, Buckley PF, Woolson S, et al. Metabolic profiles of second-generation antipyschotics in early psychosis: findings from the CAFE study. Schizophrenia Research 2009;111:9–16. [PubMed: 19398192]

[9]. Spertus J, Horvitz-Lennon M, Abing H, and Normand SL. Risk of weight gain for specific antipyschotic drugs: a Bayesian network meta-analysis of individual participant level clinical trial data. npj Schizophrenia 2018;4.

[10]. Leucht S, Samara M, Heres S, Patel MX, Woods SW, and Davis JM. Dose equivalents for second-generation antipsychotic drugs: the minimum effective dose method. Schizophrenia Bulletin 2 2014;40:314–326.

[11]. Leucht S, Samara M, Heres S, et al. Dose equivalents for second-generation antipsychotic drugs: the classical mean dose method. Schizophrenia Bulletin 6 2015;41:1397–1402.

[12]. Leucht S, Samara M, Heres S, and Davis JM. Dose equivalents for antipyschotic drugs: the DDD method. Schizophrenia Bulletin Suppl 1 2016;42:90–94.

[13]. Imai K and van Dyk DA. Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. Journal of the American Statistical Association 467 2004;99:854–866.

[14]. Gasparinni A, Armstrong B, and Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. Statistics in Medicine 29 2012;31:3821–3839.

[15]. Vehtari A, Gelman A, and Gabry J Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 5 2017;27:1413–1432.

[16]. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB. Bayesian Data Analysis, Third Edition. Boca Raton, FL: CRC Press, 2014.

[17]. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. Journal of Statistical Software 2017;76.

[18]. Burkner PC. brms: an R package for Bayesian multilevel models using Stan. Journal of Statistical Software 2017;80.

[19]. Alexander GC, Gallagher SA, Mascola A, Moloney RM, and Stafford RS. Increasing off-label use of antipyschotic medications in the United States, 1995-2008. Pharmacoepdemiol Drug Saf 2 2011;20:177–184.

[20]. Baptista T Body weight gain induced by antipsychotic drugs: mechanisms and management. Acta Psychiatrica Scandinavica 1999;100.

[21]. Nemeroff C Dosing the antipyschotic medication olanzapine. Journal of Clinical Psychiatry 1997;58.

[22]. Marder S and Meibach R Risperidone in the treatment of schizophrenia. The American Journal of Psychiatry 1994;151.

[23]. Wetterling T and Mussigbrodgt H Weight gain: side effect of atypical neuroleptics? Journal of Clinical Pyschopharmacology 1999;19.

[24]. Arvanitis L and Miller B “Seroquel” (quetiapine) in patients with acute exacerbation of schizophrenia: a comparison with haloperidol and placebo. Biological Psychiatry 1997;42.

[25]. Ciliberto N, Bossie CA, Urioste R, and Lasser RA. Lack of impact of race on the efficacy and safety of long-acting risperidone versus placebo in patients with schizophrenia or schizoaffective disorder. International Clinical Psychopharmacology 4 2005;20:207–212.

[26]. Stauffer VL, Sniadecki JL, Piezer KW, et al. Impact of race on efficacy and safety during treatment with olanzapine in schizophrenia, schizophreniform, or schizoaffective disorder. BMC Pyschiatry 89 2010;10.

[27]. Bobb JF, Dominici F, and Peng RD. Reduced hierarchical models with application to estimating health effects of simultaneous exposure to multiple pollutants. Journal of the Royal Statistical Society: Series C Applied Statistics 3 2013;62.

[28]. Normand SL. Meta-Analysis: Formulating, Evaluating, Combining, and Reporting. Statistics in Medicine 3 1999;18:321–359.

[29]. Kent DM, Steyerberg E, and van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ 2018;363.

[30]. Xie Y Population heterogeneity and causal inference. Proceedings of the National Academy of Sciences 2013;110:6262–6268.

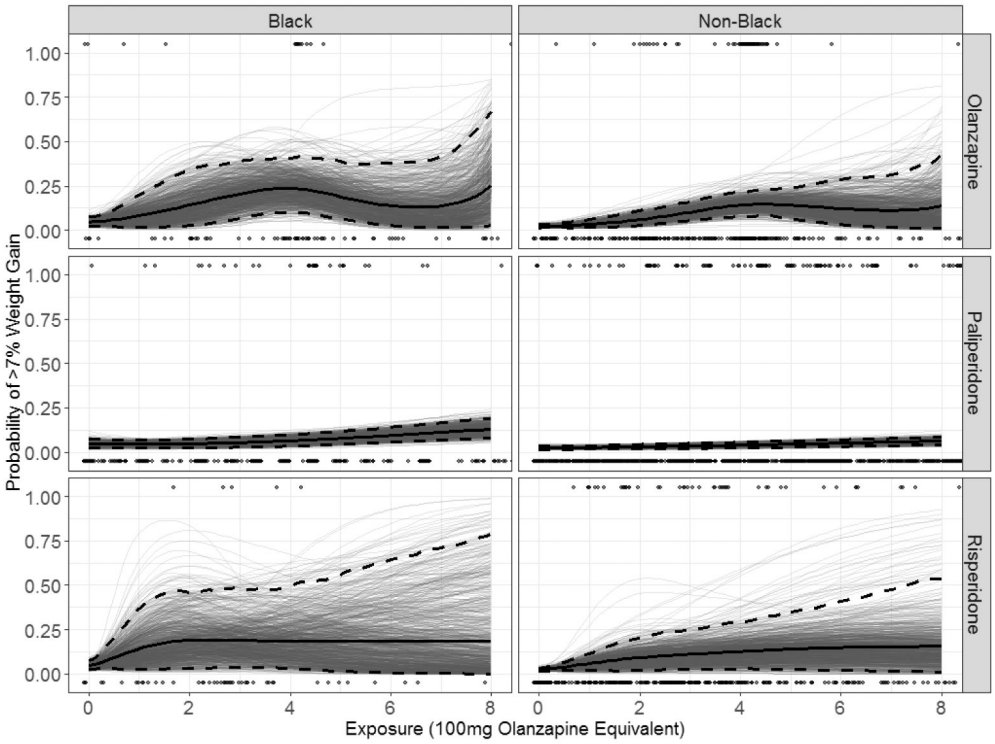

Figure 1.

Draws of estimated dose-response curves for black and non-black participants, by drug. Thin grey curves represent a single draw from the posterior, thick black curves are the posterior mean, dotted black lines bound the 95% credible region for the curves, black points above and below curves mark observed exposures and outcomes. Exposure is on the x-axis and ranges from 0 to 800mg olanzapine equivalent. The y-axis is probability of >7% weight gain. Wide credible intervals for risperidone and olanzapine reflect the fact that they were measured in very few trials and there is evidence inconsistency across trials.

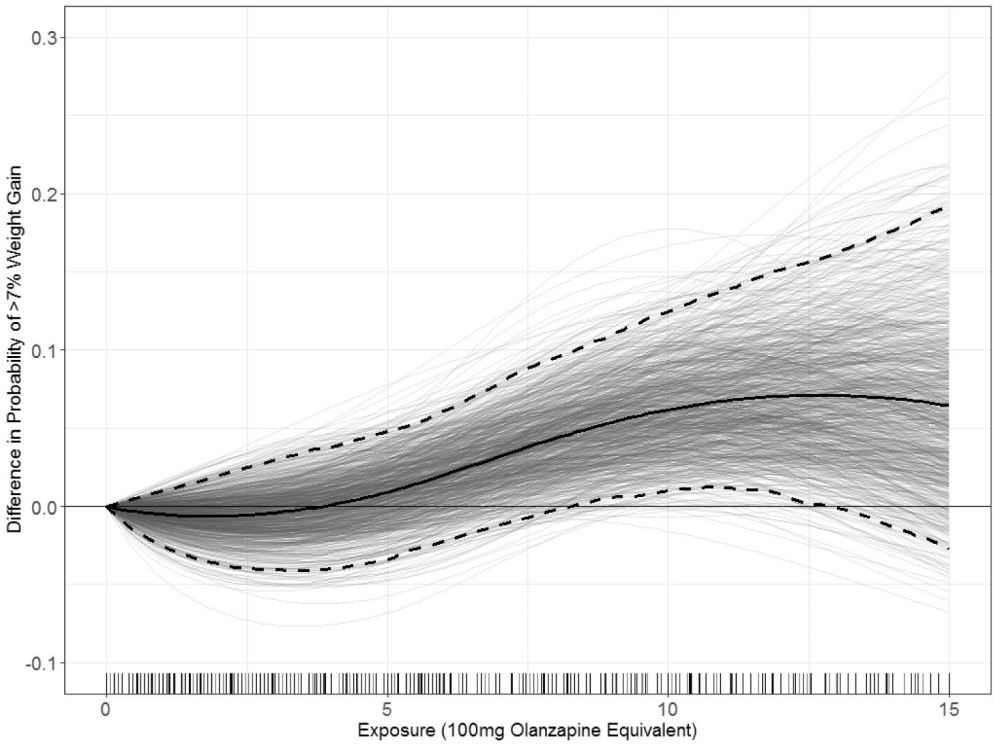

Figure 2.

Draws of estimated TEH for paliperidone showing difference between expected outcome for black and non-black participants over a range of exposures to paliperidone (x-axis). Thin grey lines are individual draws from the posterior, solid black line is posterior mean curve, and dashed black lines bound 95% credible interval. Curves above 0 indicate black participants taking paliperidone have are more likely to experience excessive weight gain compared to non-black participants for specific total cumulative dose. The y-axis is the estimated probability of weight gain for black participants minus non-black participants.

Table 1:

Characteristics of data by treatment group. %ile = percentile; OLZ = olanzapine equivalent dose; PANSS = positive and negative syndrome scale (higher scores imply more severe illness); SD = standard deviation; BMI = body mass index in kg/m^2.

	Treatment Group			
	Placebo	**Paliperidone**	**Olanzapine**	**Risperidone**
Number of Subjects (*N*)	1368	3462	527	534
Number of Trials (*J*)	12	13	4	2
Outcome and Exposure				
% > 7% Weight Gain	4.8	10.4	17.1	11.2
50th (99th) %ile 100mg OLZ	0.0 (0.0)	6.4 (26.7)	4.2 (8.4)	2.6 (20.3)
Baseline Covariates				
Mean Age (SD)	39.7 (11.9)	39.4 (11.7)	38.6 (11.1)	40.8 (11.8)
Mean PANSS (SD)	89.0 (15.1)	89.3 (13.8)	87.5 (17.0)	79.7 (14.7)
Mean BMI (SD)	27.0 (6.4)	26.9 (6.3)	27.7 (7.1)	28.7 (6.2)
% Female	37.6	38.0	31.0	34.6
% Black	18.3	17.7	15.4	9.6

Table 2:

Risk differences for percent probability of 7% weight gain under various drugs and OLZ equivalent doses. OLZ = olanzapine.

	Olanzapine	Paliperidone	Risperidone
Overall			
0 to 100mg OLZ Equivalents	1.9 (−0.8, 5.6)	0.4 (−0.4, 1.1)	5.8 (−0.5, 13.6)
0 to 500mg OLZ Equivalents	15.6 (6.7, 27.1)	3.2 (1.5, 5.2)	14.9 (0.0, 38.7)
Non-Black Participants			
0 to 100mg OLZ Equivalents	1.5 (−0.9, 4.9)	0.5 (−0.3, 1.2)	4.9 (−0.7, 12.1)
0 to 500mg OLZ Equivalents	15.2 (6.2, 26.0)	3.1 (1.4, 5.0)	14.6 (0.3, 37.7)
Black Participants			
0 to 100mg OLZ Equivalents	4.5 (−3.9, 18.6)	−0.3 (−3.1, 1.8)	11.5 (−3.0, 34.7)
0 to 500mg OLZ Equivalents	17.1 (2.2, 37.3)	4.1 (−1.2, 9.6)	16.0 (−5.3, 53.7)